Danger! Approximations...

Jean-Claude Regnier

J-C. Régnier-*Danger ! Approximations...*
pp. 99-105

Danger! Approximations...

Réflexion sur des effets d'approximations dans les calculs en statistique

Jean-Claude Régnier

Voici trois exemples dont le but est de faire comprendre l'intérêt des précautions à prendre lorsque l'on remplace une valeur par une "*valeur approchée tout à fait raisonnable*". Lorsque l'on utilise une machine à calculer, comme ceci est pratiquement une nécessité pour réaliser des calculs statistiques, il convient de travailler avec la précision maximale de l'instrument dans le déroulement des opérations. Le recours à une valeur approchée d'un ordre de grandeur compatible avec la situation d'étude n'intervient alors que pour en communiquer les informations et les conclusions. Notons cependant que l'usage de la précision maximale du calculateur automatique ne fait que diminuer le risque de provoquer les effets dont nous allons donner un exemple sans pouvoir garantir la non-réalisation de l'événement. Un tel outil ne travaille qu'avec un ensemble fini de nombres décimaux dont les nombres entiers ne sont qu'un cas particulier. Cette caractéristique est importante car elle nous rappelle quelques limites de l'instrument dans les procédures démonstratives, par exemple pour étudier le comportement d'une suite numérique déterminée par la valeur initiale : si cette valeur initiale est un nombre irrationnel comme π ou $\sqrt{2}$, il sera impossible de partir de ce point exact. Ainsi pour une suite qui serait convergente en partant de $\sqrt{2}$ et divergente à partir de toute autre valeur, l'étude empirique à l'aide d'une calculatrice s'avérera conduire inévitablement à l'erreur. Autre cas, pour comparer deux fractions $\frac{a}{b}$ et $\frac{c}{d}$ avec une calculatrice, on procède habituellement aux opérations a $\boxed{:}$ b et c $\boxed{:}$ d, certes si les affichages à l'écran sont différents, nous concluons que $\frac{a}{b} \neq \frac{c}{d}$ mais si les écritures décimales nécessitent un nombre de chiffres supérieur à la précision de l'outil et si les affichages sont identiques, nous n'avons plus qu'une présomption d'égalité car un des chiffres cachés peut être distinct.

Dans le premier exemple, nous considérons une étude portant sur le prix unitaire d'un objet en fonction de 80 points de vente existants. La variable est considérée comme une variable quantitative discrète. On cherche à obtenir le prix moyen et la dispersion des valeurs à l'aide de l'écart-type.

valeurs de la variable	6,55	6,8	6,85	6,9	7	7,1	7,35	7,4	effectif total
effectifs	7	6	8	13	16	18	7	5	80

Les calculs donnent les résultats suivants:

somme des 80 valeurs	559,4		
prix unitaire moyen	6,9925	prix unitaire moyen arrondi "raisonnablement"	7
somme des carrés des écarts à la valeur moyenne exacte	3,8005	somme des carrés des écarts à la valeur moyenne approchée	3,805
variance calculée selon la définition	0,04750625	variance calculée selon la définition	0,048
somme des carrés des valeurs de la variable	3915,405		
variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	0,04750625	**variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"**	**-0,06**

Nous repérons une aberration dans le résultat issu du calcul de la variance par la seconde méthode. En utilisant dans le cours des calculs, une valeur approchée tout à fait raisonnable (**7** pour **6,9925**), on obtient une valeur négative **- 0,06** ce qui est impossible puisque la variance est par définition positive.

Dans le second exemple, nous considérons une étude portant sur la taille d'individus La variable est considérée comme une variable quantitative continue. On cherche à obtenir la taille moyenne et la dispersion des valeurs à l'aide de l'écart-type.

valeurs de la variable	[148;152[	[152;156[	[156;160[	[160;164[	[164;168[	[168;172[	[172;176[	[176;180[	
valeurs centrales	150	154	158	162	166	170	174	178	effectif total
effectifs	5	12	21	39	33	10	2	3	125

Les calculs donnent les résultats suivants:

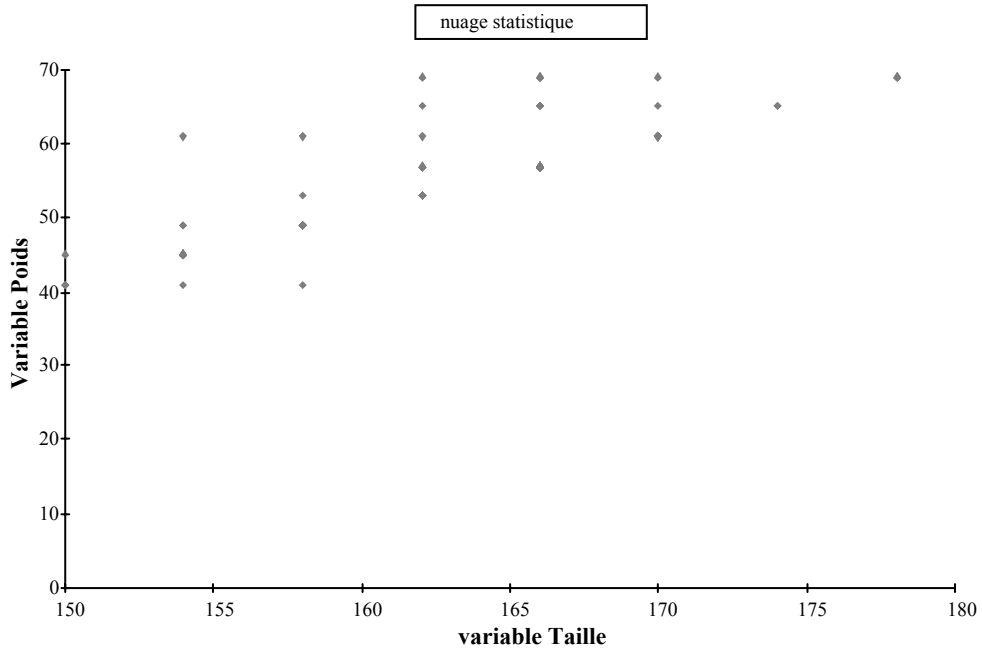
somme des 125 valeurs	20294		
TAILLE moyenne	162,352	TAILLE moyenne arrondie "raisonnablement"	162,5
somme des carrés des écarts à la valeur moyenne exacte	4032,512	somme des carrés des écarts à la valeur moyenne approchée	4035,25
variance calculée selon la définition	32,260096	variance calculée selon la définition	32,282
somme des carrés des valeurs de la variable	3298804		
variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	32,260096	**variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"**	-15,818

Nous repérons encore une aberration dans le résultat issu du calcul de la variance par la seconde méthode. En utilisant dans le cours des calculs, une valeur approchée tout à fait raisonnable (**162,5** pour **162,352**), on obtient ici encore une valeur négative **-15,818**

Dans ce troisième exemple, nous allons considérer le croisement de la variable *Taille des individus* avec la variable *Poids des individus*. Le tableau statistique ci-après fournit les résultats à l'issue d'une enquête auprès de 125 individus.

valeurs de la variable	Taille (cm) X	[148;152[	[152;156[	[156;160[	[160;164[	[164;168[	[168;172[	[172;176[	[176;180[	
Poids (kg) Y	*valeurs centrales*	150	154	158	162	166	170	174	178	effectif
[39;43[	41	4	1	1	0	0	0	0	0	6
[43;47[	45	1	8	0	0	0	0	0	0	9
[47;51[	49	0	2	18	0	0	0	0	0	20
[51;55[	53	0	0	1	32	0	0	0	0	33
[55;59[	57	0	0	0	4	26	0	0	0	30
[59;63[	61	0	1	1	1	0	8	0	0	11
[63;67[	65	0	0	0	1	4	1	2	0	8
[67;71[	69	0	0	0	1	3	1	0	3	8
	effectifs	5	12	21	39	33	10	2	3	125

Ci-après, la représentation graphique du nuage des 125 points apporte des informations sur une éventuelle liaison statistique entre ces deux variables biométriques.



Cette recherche de lien est aussi étayée par le recours au coefficient de corrélation linéaire de Bravais-Pearson R_{BP} entre deux variables quantitatives X et Y. Nous donnons l'expression algébrique :

$R^2_{BP} = \dfrac{Cov^2(X;Y)}{V(X)\,V(Y)}$ où Cov(X ; Y) désigne la covariance de X et de Y , V(X) et V(Y) la variance respective de X et de Y. Par construction les valeurs de R^2_{BP} sont comprises entre 0 et 1 donc celles de R_{BP} , entre -1 et 1.

Nous convenons de calculer leurs valeurs par les algorithmes suivant :

Variance = moyenne des carrés des valeurs de la variable - carré de la moyenne

Covariance = moyenne des produits des valeurs des deux variables - produit des deux moyennes

Le tableau ci-dessous donne alors les valeurs obtenues à partir de diverses conduites d'approximation :

	résultats exacts	*résultats approchés (1)*	*résultats approchés (2)*
somme des valeurs de X	20294	***	***
moyenne de X	**162,352**	**162,4**	**162,3**
somme des carrés de X	3298804	***	***
moyenne des carrés de X	26390,432	***	***
carré de la moyenne de X	26358,171904	26373,76	26341,29
somme des valeurs de Y	6833	***	***
moyenne de Y	**54,664**	**54,7**	**54,7**
somme des carrés de Y	379317	***	***
moyenne des carrés de Y	3034,536	***	***
carré de la moyenne de Y	2988,152896	2992,09	2992,09
somme des produits XY	1113454	***	***
moyenne des produits XY	8907,632	***	***
produit des moyennes	8874,809728	8883,28	8877,81
covariance cov(X,Y)=	32,8222719999994	24,351999999999	29,8219999999983
$cov^2(X;Y)$ =	1077,30153924195	593,019903999949	889,351683999898
variance V(X)=	32,2600959999982	16,6719999999987	49,1419999999962
variance V(Y)=	46,383104	42,4459999999999	42,4459999999999
coefficient $R^2_{BP}(X,Y)$=	0,719965715975999	0,838001505446728	0,426367344276117

	résultats approchés (3)	*résultats approchés (4)*	*résultats approchés (5)*
somme des valeurs de X	***	***	***
moyenne de X	**162,4**	**162,5**	**162**
somme des carrés de X	***	***	***
moyenne des carrés de X	***	***	***
carré de la moyenne de X	26373,76	26406,25	26244
somme des valeurs de Y	***	***	***
moyenne de Y	**54,6**	**54,66**	**55**
somme des carrés de Y	***	***	***
moyenne des carrés de Y	***	***	***
carré de la moyenne de Y	2981,16	2987,7156	3025
somme des produits XY	***	***	***
moyenne des produits XY	***	***	***
produit des moyennes	8867,04	8882,25	8910
covariance cov(X,Y)=	40,5919999999987	25,3819999999996	-2,36800000000039
$cov^2(X;Y)$ =	1647,7104639999	644,24592399998	5,60742400000186
variance V(X)=	16,6719999999987	-15,8179999999993	146,432000000001
variance V(Y)=	53,3759999999997	46,8204000000005	9,53600000000006
coefficient $R^2_{BP}(X,Y)$=	1,85160000598374	-0,869891302577595	0,004015699066036

Dans la colonne *Résultats exacts,* nous rapportons effectivement les valeurs exactes obtenues à partir des calculs. Dans les autres colonnes nous avons adopté une conduite d'approximation en modifiant les valeurs des moyennes en choisissant des *valeurs concrètement raisonnables.* Force est de constater les effets spectaculaires que ces choix peuvent provoquer.

	résultats exacts	*résultats approchés (1)*	*résultats approchés (2)*
moyenne de X	**162,352**	**162,4**	**162,3**
moyenne de Y	**54,664**	**54,7**	**54,7**
coefficient $R^2_{BP}(X,Y)$=	0,719965715975999	0,838001505446728	0,426367344276117
coefficient de Bravais-Pearson $R_{BP}(X,Y) \approx$	0,8485	0,9154	0,6529

	résultats approchés(3)	*résultats approchés(4)*	*résultats approchés (5)*
moyenne de X	**162,4**	**162,5**	**162**
moyenne de Y	**54,6**	**54,66**	**55**
coefficient $R^2_{BP}(X,Y)=$	1,85160000598374	-0,869891302577595	0,004015699066036
coefficient de Bravais-Pearson $R_{BP}(X,Y) \approx$	absurde	absurde	0,0633

Comment ne pas s'étonner devant les faits suivants :

- en arrondissant le poids moyen de 54,664 kg à 54,7 kg et la taille moyenne de 162,352 cm à la valeur au dixième près par excès c'est à dire en prenant 162,4 cm ou à la valeur au dixième près par défaut c'est à dire en prenant 162,3 cm , nous faisons passer la valeur du coefficient $R^2_{BP}(X,Y) \approx 0,7199$ de 0,8300 à 0,4236 ce qui revient après l'avoir augmenté à le diminuer de moitié !

- en choisissant les valeurs approchées au dixième près suivantes, par excès pour la taille moyenne avec 162,4 cm et par défaut pour le poids moyen avec 54,6 kg, nous obtenons une aberration avec une valeur supérieure à 1!

- en choisissant une valeur approchée au 1/100 près par défaut pour le poids moyen, c'est à dire 54,66 kg et la valeur centrale 162,5 cm pour la taille moyenne, nous obtenons une valeur aberrante parce que négative !

- en choisissant d'arrondir aux unités ce qui revient à considérer un poids moyen de 55 kg et une taille moyenne de 162 cm, nous aboutissons à une valeur coefficient $R^2_{BP}(X,Y)$ proche de 0 !

De cela vont résulter des décisions opposées. Ainsi une valeur du coefficient $R^2_{BP}(X,Y)$ voisine de 0,8380, c'est à dire une valeur du coefficient de Bravais-Pearson voisine de 0,91, incite à rejeter l'hypothèse d'absence de lien linéaire. Mais il n'en est pas de même avec une valeur de $R^2_{BP}(X,Y)$ voisine de 0,0040 qui donne une valeur de $R_{BP}(X,Y) \approx 0,06$. Cette dernière incite à renoncer à ce rejet par référence au cadre théorique qui donne un sens à ce coefficient et dont nous ne faisons pas état ici.

En ce qui concerne la situation où le coefficient prend une valeur supérieure à 1, il m'a été donné de l'observer effectivement lors d'une surveillance d'épreuve de baccalauréat , il y a une dizaine d'années. J'ai pu alors me rendre compte par hasard qu'une candidate se débattait avec sa

calculatrice et que ses résultats l'étonnaient. Recommençant plusieurs fois, elle trouvait cependant toujours un coefficient supérieur à 1. Force est de constater qu'elle entrait à chaque fois des valeurs arrondies qui la plaçait dans des conditions analogues à celles que j'ai tentées de décrire dans cet article. Les erreurs ainsi produites ne relèvent pas d'une incompréhension des notions statistiques mises en œuvre mais bien des limites posées par un instrument de calcul. L'enseignement a tout intérêt à intégrer ce genre de problème dans l'éducation statistique des élèves. Le but principal de cet article est d'attirer l'attention des enseignants et des étudiants sur ce phénomène, effet secondaire des avantages majeurs qu'apporte les calculateurs automatiques.